EXTENDED REPORT

Development and preselection of criteria for short term improvement after anti-TNFα treatment in ankylosing spondylitis

J Brandt, J Listing, J Sieper, M Rudwaleit, D van der Heijde, J Braun

Ann Rheum Dis 2004;**63**:1438–1444. doi: 10.1136/ard.2003.016717

See end of article for authors' affiliations

Correspondence to: Professor J Braun, Rheumazentrum Ruhrgebiet, Landgrafenstr 15, 44652 Herne, Germany; j.braun@rheumazentrum-ruhrgebiet.de

Accepted 24 January 2004
Published online first 25 March 2004

Objective: To develop and compare candidate improvement criteria for anti-TNFα treatment in ankylosing spondylitis with optimal discriminating capacity between treatment and placebo.
Methods: Data from two randomised controlled trials which included 99 patients treated with infliximab or etanercept were used to evaluate 50 candidate improvement criteria. These were developed on the basis of pain, patient's global assessment, function, morning stiffness, spinal mobility, and C reactive protein. Different levels of improvement in each domain (20–60%) were used to define Boolean type criteria. These criteria were compared with different percentages of improvement on the BASDAI and with modified ASAS improvement criteria. Bootstrap methods were applied to calculate 95% confidence intervals (CI) of the χ^2 test values to select the best candidate improvement criteria.
Results: The best performing improvement criteria were ''20% improvement in five of six domains'' ($\chi^2 = 31.9$ (95% CI, 18.0 to 46.9)) with a low placebo response of 2.9% and a high response to infliximab of 67.7%; and ''ASAS 40% improvement'' ($\chi^2 = 26.5$ (13.3 to 41.1)), with response to placebo of 5.7% and response to infliximab of 64.7%. The good discriminating capacity of the two improvement criteria was confirmed by the combined dataset of the infliximab and etanercept trial.
Conclusions: The ''five of six'' improvement criterion has the advantage of including the objective domains *spinal mobility* and *acute phase reactants*, but requires only 20% improvement. The ASAS 40% improvement criterion has the advantage of setting a high threshold, but only in patient reported outcomes. The choice between these improvement criteria needs to be based on further validation from upcoming trials.

Several recent trials with the anti-tumour necrosis factor α (anti-TNFα) agents infliximab[1–5] and etanercept[6–8]—both of open label type and double blind, randomised, controlled—have suggested that these agents are effective in active ankylosing spondylitis (AS). TNFα blocking agents are generally considered to be a major breakthrough in the overall treatment of this disease. In contrast to rheumatoid arthritis, Crohn's disease, and psoriatic arthritis—in which corticosteroids and disease modifying antirheumatic drugs (DMARD) have established efficacy—such treatment has only limited value in AS.[9] In fact, at present the treatment of AS consists mainly of physiotherapy, non-steroidal anti-inflammatory drugs (NSAIDs), local corticosteroid injections, and DMARD therapy with sulphasalazine in patients with peripheral arthritis.

When the first randomised trial with infliximab[1] was designed in 2000, there were no established improvement criteria in AS at all. From our pilot study we knew about the strong efficacy of anti-TNFα therapy with infliximab in active AS.[10 11] We therefore chose 50% improvement in the Bath ankylosing spondylitis disease activity index (BASDAI)[12] as the major outcome and improvement criterion, in analogy to the American College of Rheumatology (ACR) 50% criterion.[13 14] This was also used in the subsequent German randomised controlled trial (RCT)[1] and it has frequently been adopted in later trials by other investigators.[2–7] The consensus conference on anti-TNFα treatment in AS which took place in January 2003 in Berlin included 50% improvement of BASDAI in its recommendations for discontinuation of anti-TNFα treatment in clinical practice.[15]

The Assessments in Ankylosing Spondylitis (ASAS) working group published a preliminary definition of short term improvement in AS based on the best discrimination between NSAID treatment (50%) and placebo (25%).[16] These ASAS criteria are based on improvement of at least 20% and 1 unit (on a 0–10 scale) in three of the following four domains: patient's global assessment, pain, function, and morning stiffness. In addition, it is required that there is no worsening by more than 20% in the remaining domain. It was unclear whether the ASAS improvement criteria were also be the best criteria to assess the efficacy of anti-TNFα treatment. There are some reasons to consider why this might not be so. Initial NSAID trials are always based on the flare design: patients who are in need of NSAIDs are asked to stop the drug they are taking and only if they show a flare they can be randomised. The trial design in the anti-TNFα studies has been completely different. Patients with very active disease despite optimal doses of NSAIDs are included in the trial. Even in this group of patients with very severe disease, unresponsive to NSAIDs, there is a major treatment response of anti-TNFα therapy. The difference in trial design might also have implications for the placebo response and the regression to the mean effect.

Abbreviations: ACR, American College of Rheumatology; AS, ankylosing spondylitis; ASAS, Assessments in Ankylosing Spondylitis working group; BASDAI, Bath ankylosing spondylitis disease activity index; BASFI, Bath ankylosing spondylitis functional index; BASMI, Bath ankylosing spondylitis metrology index; DCART, disease controlling anti-rheumatic therapy; DMARD, disease modifying antirheumatic drug; NSAID, non-steroidal anti-inflammatory drug; RCT, randomised controlled trial; SMARD, symptom modifying anti-rheumatic drug; SRM, standardised response mean; TNFα, tumour necrosis factor α; VAS, visual analogue scale

Table 1 Patient characteristics in (A) the infliximab trial and (B) the etanercept trial

Characteristic	Active drug	Placebo
(A) Infliximab trial		
n	34	35
Male/female ratio	23/11	22/13
Age (years) (mean (SD))	40.6 (8.0)	39.0 (9.1)
Disease duration (years) (mean (SD))	16.4 (8.3)	14.9 (9.3)
HLA-B27 positive (%)	31 (91.2)	27 (87.5)
Number of swollen joints, range 0–68 (5% trimmed mean (SD))	0.9 (4.1)	1.3 (5.2)
Number of enthesitic regions, range 0–12 (mean (SD))	1.7 (3.3)	2.0 (3.2)
History of anterior uveitis (%)	17 (50)	15 (43)
BASDAI (mean (SD))	6.5 (1.2)	6.3 (1.4)
BASFI (mean (SD))	5.4 (1.8)	5.1 (2.2)
BASMI (mean (SD))	3.7 (2.0)	3.7 (2.2)
Pain (VAS) (mean (SD))	7.2 (1.6)	7.3 (1.7)
Radiological score for spine (BASRI-s) (mean (SD))	6.5 (2.5)	6.6 (2.9)
(B) Etanercept trial		
n	14	16
Male/female ratio	10/4	12/4
Age* (years) (mean (SD))	39.8 (9.1)	32.0 (7.5)
Disease duration (years) (mean (SD))	14.9 (8.3)	11.4 (8.8)
HLA-B27 positive (%)	12 (85.7)	15 (93.8)
Number of swollen joints, range 0–68 (mean (SD) at baseline)	0.9 (1.5)	1.7 (4.0)
Number of enthesitic regions, range 0–12 (mean (SD) at baseline)	1.4 (2.2)	1.3 (1.7)
History of anterior uveitis (n (%))	5 (35.7)	3 (18.8)
BASDAI (mean (SD))	6.5 (1.2)	6.6 (1.0)
BASFI (mean (SD))	6.2 (1.8)	5.3 (2.3)
BASMI (mean (SD))	4.1 (1.7)	3.8 (2.1)
Pain (VAS) (mean (SD))	7.4 (1.8)	7.6 (1.2)
Radiological score for spine (BASRI-s) (mean (SD))	6.3 (2.5)	5.4 (1.7)

In the infliximab trial, the baseline characteristics of the two treatment groups showed no significant differences. *$p<0.05$ calculated by the Wilcoxon rank sum test between groups; all other baseline characteristics of the groups in the etanercept trial showed no significant differences.
BASDAI, Bath ankylosing spondylitis disease activity index; BASFI, Bath ankylosing spondylitis functional index; BASMI, Bath ankylosing spondylitis metrology index; BASRI-s, Bath ankylosing spondylitis radiology index of the spine; VAS, visual analogue scale.

Another issue is the choice of domain. In NSAID trials the domain *spinal mobility*, which seems rather important for patients, proved unresponsive to treatment, and the same was true in the acute phase reactant trials. Anti-TNFα therapy has a positive effect on both domains, which are part of the ASAS core set to assess treatment with a presumed DCART (disease controlling anti-rheumatic therapy) effect. Therefore it was necessary to repeat the process of developing improvement criteria for this specific aim and to compare the novel criteria with the ASAS 20 and the often used BASDAI 50.

The domains used for the improvement criteria are based on the core set of outcome measures recently proposed by the ASAS working group[17 18] including validated measures of function,[19] spinal mobility,[20] and acute phase reactants. This study provides data that will enable us to choose the best set of improvement criteria to assess changes in AS.

METHODS

Choice of domains for outcome measurement in AS

Six different outcome domains were chosen. The first four are already included in the ASAS improvement criteria, which were derived from the ASAS core set: pain, patient's global assessment, function, and morning stiffness. These domains were assessed by a visual analogue scale (VAS) for pain and patient global assessment, by the Bath ankylosing spondylitis functional index (BASFI)[19] for function, and by the two last questions of the BASDAI for morning stiffness.[12] Two additional domains were added in this analysis: spinal mobility and acute phase reactants. These are also included in the core set for DCART.[17] The tools that were used to assess these were the Bath ankylosing spondylitis metrology index (BASMI)[20] and C reactive protein. The BASMI comprises semiquantitative assessments of anterior lumbar flexion (modified Schober), lateral lumbar flexion, cervical rotation, occiput to wall distance, and intermalleolar distance.

On the basis of these six domains, many different candidate criteria for anti-TNFα treatment in AS were developed. The first set of criteria was based on the primary list of all six domains, with the following modifications:

- The domains *pain* and *morning stiffness* were replaced by the BASDAI,[12] the most frequently used measure for disease activity in AS, which includes both these domains but also assesses other important features of AS such as peripheral arthritis, enthesitis, and fatigue.
- The domain *acute phase reactants* was excluded because it is not entirely clear if acute phase reactants reflect disease activity reliably, and they are not increased in all AS patients.[21]
- The domain *spinal mobility* was assessed by the single item *examination of lateral lumbar flexion* instead of the whole BASMI. This was because lateral lumbar flexion alone had the highest standardised response mean (SRM) among all five individual BASMI items. Thus we assumed this combined measure performed as well as the total BASMI (see below) but was easier to do. *Chest expansion* was not assessed in the trials.

Ways of defining improvement

The different sets of domains were combined to form Boolean-type improvement criteria. Such criteria require improvement in all or at least a specified subset of the domains at a certain specified level. ACR response criteria for rheumatoid arthritis are an example of Boolean-type improvement criteria. They include criteria requiring improvement in all domains or in, for example, four of six

Table 2 Development of subset assessment of candidate improvement criteria for anti-TNFα agents in ankylosing spondylitis on the basis of the six domains of pain, patient global, function, inflammation, spinal mobility, and C reactive protein

		Infliximab trial (n = 69)				
Criterion	Improvement definition	Per cent improving in placebo group (n = 35)	Per cent improving in infliximab group (n = 34)	χ^2	95% CI of the χ^2 value	Combined dataset of etanercept and infliximab trials (n = 69) (χ^2)
Relative improvement						
1	**≥20% change in any five of six domains**	**2.86**	**67.65**	**31.91***	**18.02 to 46.94**	**27.43**
2	≥30% change in any five of six	2.86	50.00	19.88	9.05 to 32.53	15.89
3	≥40% change in any five of six	2.86	44.12	16.48	6.99 to 27.81	15.95
4	≥30% change in any four of six	5.71	64.71	26.46*	13.86 to 41.11	20.44
5	≥40% change in any four of six	5.71	61.76	24.38*	11.52 to 39.20	23.09
6	≥50% change in any four of six	5.71	52.94	18.69	7.93 to 32.38	19.48
7	≥30% change in any three of six	22.86†	73.53	17.75	4.96 to 34.02	25.88
8	≥40% change in any three of six	8.57	67.65	25.63*	12.72 to 40.36	22.35
9	≥50% change in any three of six	8.57	61.76	21.51*	9.51 to 36.91	21.95
Relative and absolute improvement						
10	≥20%/10 units change in any five of six domains	2.86	64.71	29.69*	16.26 to 44.25	27.72
11	≥30%/10 units change in any five of six	2.86	47.06	18.15	7.90 to 30.34	15.89
12	≥40%/20 units change in any five of six	0.0	38.24	16.49	8.53 to 26.67	15.78
13	≥30%/10 units change in any four of six	5.71	64.71	26.46*	13.86 to 41.11	20.84
14	≥40%/20 units change in any three of six	8.57	67.65	25.63*	12.72 to 40.36	22.64
15	≥50%/20 units change in any three of six	8.57	58.82	19.60*	8.32 to 33.75	21.95
16	≥60%/30 units change in any three of six	0.0	47.06	21.44*	11.91 to 32.53	17.04
Relative improvement including pain						
17	≥20% change in any five of six domains, including pain	2.86	67.65	31.91*	18.02 to 46.94	27.43
18	≥40% change in any four of six, including pain	5.71	58.82	22.40*	10.14 to 36.96	23.09
Combinations in relative improvement						
19	≥40% in four of six domains or ≥30% in five of six	5.71	61.76	24.38*	11.52 to 39.20	20.69
20	≥50% in four of six or ≥30% in five of six	5.71	58.82	22.40*	10.14 to 36.71	20.69

Columns 3–7: infliximab trial, week 12. Last column: data from the combined dataset at week 6; to be comparable, n = 69 from n = 99 at random. The criteria that performed best are shown in bold.
*Candidate criteria with χ^2 values that are not statistically significant below reference criterion 1.
†Placebo response significantly above 10%.
CI, confidence interval.

domains. Finally, we defined additional candidate improvement criteria with mandatory improvement in one of the domains. For this, we selected three domains of high clinical relevance for the evaluation of efficacy of anti-TNFα treatment: pain, function, and spinal mobility (for example, a change in any five of six domains would include a mandatory improvement in spinal mobility). We show only selected results of all the candidate improvement criteria investigated.

Improvement in a single domain was determined by a combination of relative and absolute change from baseline values or by a relative change alone. Relative changes at levels of ≥20%, ≥30%, ≥40%, ≥50%, and ≥60% improvement were chosen. If relative and absolute changes were combined, the following combinations were applied: ≥20% change plus 1 unit on a 0–10 scale, ≥30% change plus 1 unit, ≥40% change plus 2 units, ≥50% change plus 2 units, and ≥60% change plus 3 units, respectively.

In some of the recently published trials on new treatments for AS, response was defined as improvement in the single domain *disease activity* as assessed by the BASDAI using cut off values between 20% and 70%. In order to compare these with the multiple domain improvement criteria defined above, we calculated the 20% to 70% improvements in the BASDAI response in steps of 10%. Additionally, we calculated the ASAS improvement criteria for several cut off values ranging from 30% to 70% for further comparison. For this, the original ASAS 20 improvement criteria were modified in the following ways: an ASAS 40%, 50%, 60%, and 70% response was defined as a relative improvement in three of four domains, with an absolute improvement of at least 2 units (3 units) and no deterioration in the remaining domain.

Patients

The first dataset to evaluate candidate improvement criteria contained data from week 12 of the placebo controlled part of the ''infliximab in AS'' trial with 69 active AS patients,[1] who were treated with either 5 mg/kg infliximab or placebo in weeks 0, 2, and 6. The discriminant properties of the candidate improvement criteria were validated with a second dataset which included results from the ''etanercept in AS'' trial.[8] In this trial patients received 25 mg etanercept or placebo subcutaneously twice weekly. The clinical efficacy of infliximab and etanercept was rather similar. In the etanercept trial fewer patients were included (n = 30) and the placebo controlled phase was only six weeks long, as compared with 12 weeks in the infliximab trial. In both trials the patients were selected using the same inclusion and

Table 3 Development of subset assessment of candidate improvement criteria for anti-TNFα agents in ankylosing spondylitis on the basis of the six domains: pain, patient global, function, inflammation, spinal mobility, and C reactive protein, modified by exchange of certain domains and instruments

Criterion	Improvement definition	Infliximab trial (n = 69)				Combined dataset etanercept and infliximab trial (n = 69) (χ^2)
		Per cent improving in the placebo treated group (n = 35)	Per cent improving in the infliximab treated group (n = 34)	χ^2	95% CI of the χ^2 value	
Five domains including the BASDAI instead of pain and inflammation						
21	≥20%/10 units change in any four of five domains	5.71	67.64	28.63*	4.25 to 18.69	22.04
22	≥20% change in any four of five	5.71	67.65	28.63*	14.34 to 44.36	23.22
23	≥20% change in any four of five, including BASDAI	5.71	67.64	28.63*	14.34 to 44.36	23.22
24	≥40%/20 units change in any three of five	5.71	58.82	22.40*	9.05 to 32.53	24.72
25	≥40% change in any three of five	5.71	64.71	26.46*	12.97 to 40.72	27.40
26	≥40% in 3 of 5 or ≥30% in four of five, including BASDAI	2.86	64.71	29.69*	16.61 to 43.92	23.43
Five domains for improvement in AS without the domain acute phase reactants						
27	≥20%/10 units change in any four of five domains	8.57	73.52	30.18*	16.13 to 46.18	25.88
28	≥20% change in any four of five	8.57	76.47	32.63*	17.26 to 48.41	25.43
29	≥20% change in any four of five, including pain	8.57	76.47	32.63*	17.26 to 48.41	25.43
30	≥40%/20 unit change in any three of five domains	8.57	64.71	23.52*	11.25 to 37.87	18.69
31	≥40% change in any three of five	8.57	67.65	25.63*	12.72 to 40.36	20.96
32	≥40% in three of five or ≥30% in four of five, including pain	8.57	64.71	23.52*	10.88 to 38.20	19.80
Six domains for improvement in AS with an exchange of BASMI for lateral lumbar flexion						
33	≥20%/10 unit change in any five of six domains	0.0	58.82	28.99*	17.81 to 42.36	25.62
34	≥30%/10 unit change in any four of six	5.71	64.71	26.46*	13.86 to 41.11	20.84
35	≥40%/20 unit change in any three of six	8.57	67.65	25.63*	12.72 to 40.36	19.94
36	≥20% change in any five of six	14.29	76.47	26.95*	13.22 to 43.41	28.03
37	≥30% change in any four of six	5.71	61.76	24.38*	12.27 to 38.43	22.33
38	≥40% change in any three of six	11.43	70.59	25.03*	11.79 to 40.52	29.24

Columns 3–7: infliximab trial, week 12. Last column: data from the combined dataset at week 6; to be comparable, n = 69 from n = 98 at random.
*Candidate criteria with χ^2 values that are not significantly below reference criterion 1 (table 2).
AS, ankylosing spondylitis; BASMI, Bath ankylosing spondylitis metrology index; CI, confidence interval.

exclusion criteria (table 1). Only patients with high disease activity—defined by a BASDAI value of at least 4 and pain of at least 4 on a VAS despite treatment with NSAIDs—were included in both studies, in order to ensure that these patients were suitable candidates for treatment with the anti-TNFα agents. The patients in both trials were in a very active state of disease, with mean BASDAI values at baseline of around 6.5 in both trials.

Statistical evaluation

A selection was made to determine the best performing newly developed improvement criteria using the data from the infliximab trial, based on the following rules.

(1) Criteria with a placebo response above 10% were excluded. This cut off was chosen arbitrarily by expert opinion in order to omit improvement criteria with a high placebo response from further evaluation and to select for other improvement criteria with a high specificity to identify patients who respond to anti-TNFα therapy. This rule was determined by calculating 95% confidence intervals of the placebo response rates. Thus improvement criteria with a significantly higher placebo response than 10% could be identified.

(2) Criteria with a low placebo response (below 10%) and a high power to detect the difference between placebo and effective treatment were considered further. The power of the different criteria was evaluated by comparing their χ^2 test values. By bootstrapping analysis 95% confidence intervals of these χ^2 values were calculated. As there is no exact subset selection procedure available for dependent χ^2 values, an approximate bootstrap procedure was applied to investigate which criteria were significantly different from those with the highest χ^2. For that purpose all improvement criteria not already excluded because of the high placebo response were compared with the improvement definition that had the highest power (highest χ^2 value; reference criteria). All criteria with a significantly lower power than the reference criteria were excluded.

A validation step was carried out by comparing the 12 week results of the infliximab trial with a combined sample of six week data from the infliximab and etanercept trials. In four patients withdrawn from the trial before week 12 the ''last observation carried forward'' method was applied to estimate the 12 week data by their six week data.

RESULTS

Candidate improvement criteria were evaluated using the data from the infliximab trial.[1] Most of these were very strict—only four of 50 candidate improvement criteria shown in tables 2 to 4 had placebo response rates that were significantly higher than 10% (Nos 7, 39, 40, and 45). These criteria were removed from further consideration.

Of the remaining criteria, No 1 (table 2) performed best among those that were based on the original six domains ($\chi^2 = 31.9$ (95% CI, 18 to 47)). This high discriminative power is based on a very low placebo response of 2.9% and a high response in the group treated with infliximab of 67.7%. It was

Table 4 Improvement definitions by the single domain disease activity measured by the Bath ankylosing spondylitis disease activity index and by the criteria of the Assessment in Ankylosing Spondylitis working group on different improvement levels

Criterion	Improvement definition	Infliximab trial: Per cent improving in the placebo treated group (n = 35)	Infliximab trial: Per cent improving in the infliximab treated group (n = 34)	Infliximab trial: χ^2	Infliximab trial: 95% CI of the χ^2 value	Combined dataset etanercept and infliximab trial (n = 69) (χ^2)
Relative improvement of disease activity						
39	≥20% change of the BASDAI	37.14†	85.29	16.79	4.72 to 32.04	12.52
40	≥30% change of the BASDAI	25.71†	73.53	15.78	3.68 to 30.83	17.10
41	≥40% change of the BASDAI	8.57	64.71	23.52*	10.16 to 39.30	21.31
42	≥50% change of the BASDAI	8.57	58.82	19.60*	7.46 to 34.04	20.31
43	≥60% change of the BASDAI	5.71	38.24	10.72	2.45 to 22.19	15.67
44	≥70% change of the BASDAI	5.71	26.47	5.54	0.30 to 15.41	9.87
ASAS improvement criteria						
45	≥20% and 10 units	25.71†	73.53	15.78	4.41 to 30.72	23.66
46	≥30% and 10 units	11.43	64.71	20.85*	8.59 to 36.91	20.19
47	**≥40% and 20 units**	**5.71**	**64.71**	**26.46***	**13.30 to 41.14**	**17.34**
48	≥50% and 20 units	5.71	52.94	18.69	7.74 to 32.39	15.01
49	≥60% and 30 units	0.0	44.12	19.73	10.48 to 30.45	13.88
50	≥70% and 30 units	0.0	32.35	13.47	6.07 to 22.56	5.38

Columns 3–7: infliximab trial, week 12. Last column: data from the combined dataset at week 6; to be comparable, n = 69 from n = 98 at random. The criterion which performed best is shown in bold.
*Candidate criteria with χ^2 values that are not significantly below reference criterion 1 (table 3).
†Placebo response significantly above 10%.
ASAS, Assessment in Ankylosing Spondylitis working group; BASDAI, Bath ankylosing spondylitis disease activity index; CI, confidence interval.

used as the reference criterion to investigate which of the others differed significantly from improvement criterion No 1.

There were two criteria with slightly higher χ^2 values, but these did not include the original set of six domains and did not differ significantly from criterion No 1, so they were not considered as reference improvement criteria. In detail, these two improvement criteria combined five domains because the domain *acute phase reactants* was omitted. The first criterion (No 28, table 3) defined improvement by a ≥20% change in four of five domains ($\chi^2 = 32.6$) and the second (No 29) was a modification of the first by adding a mandatory improvement of pain (≥20% change in four of five domains including pain; $\chi^2 = 32.6$). Furthermore, candidate criteria were omitted from analysis when their χ^2 values were significantly lower than reference criterion 1 (table 2) by bootstrap analysis. This resulted in 36 improvement criteria with a high discriminative power to detect differences between placebo and treatment with infliximab, which are marked by an asterisk in tables 2, 3, and 4.

In more detail, the following conclusions can be drawn from the calculations done with these 36 improvement criteria using the data from the infliximab RCT.

Improvement criteria Nos 4, 5, and 8 (table 2)—which defined improvement in at least three or four of six domains—had a similar power for discriminating between the active drug and placebo as reference criterion No 1. However, the results of these improvement criteria mainly reflected the original ASAS domains (pain, patient global, function, inflammation) and not the more objective measures of C reactive protein and spinal mobility. For example, 75% of the responders of improvement criterion No 4 responded in all of the four original domains and only 42% in C reactive protein or spinal mobility. For improvement criterion No 5 the figures were 65% and 39%, respectively.

Improvement criteria which defined improvement only by a relative change—for example, criterion 1 (table 2)—performed as well as the corresponding improvement criteria where relative and absolute improvements were combined (as in criterion 10 (table 2)). This was mainly because the baseline values were high and consequently a (large) relative improvement immediately results in a corresponding (high) absolute value. Based on feasibility, improvement criteria with definitions dealing only with relative change are preferable to improvement criteria reflecting relative and absolute change.

Improvement criteria with combinations of different values of relative improvement—for example, criterion 19 (table 2) with ≥40% in four of six domains or ≥30% in five of six domains—and those with combinations of different values of relative improvement and absolute improvement showed very similar results to the corresponding more feasible improvement criteria (Nos 2 and 5). Moreover, they showed somewhat lower responses in the group treated with the active drug than were found with reference criterion 1, so these improvement criteria performed worse for defining response.

We also analysed another subset of candidate improvement criteria with mandatory improvement in a specific domain that was considered to be highly relevant clinically (pain, function, and spinal mobility). Improvement criteria with a mandatory improvement in pain (Nos 17 and 18, table 2) had a similarly good performance to those without this restriction (Nos 1 and 5, table 2). Thus a mandatory improvement in pain is not necessary for a definition of improvement with anti-TNFα drugs. Improvement criteria with a necessary improvement in domain function and spinal mobility showed a worse performance, with significant differences in their χ^2 values compared with reference criterion 1 (data not shown), indicating that a mandatory improvement in function or spinal mobility had less statistical power.

With differently modified improvement criteria—where disease activity was assessed using the BASDAI instead of the domains *pain* and *morning stiffness*, resulting in improvement criteria with five domains (table 3)—the performance was nearly as good as with six domains (table 2). In detail, criterion 23 in table 3, with a definition of ≥20% change in any four of five domains including the BASDAI, had a slightly less good χ^2 value (28.6) than the reference improvement criterion of ≥20% change in any five of six domains ($\chi^2 = 31.9$). Because these differences might be from chance alone, it can be assumed that criteria with inclusion of the BASDAI had a similar good performance to criteria with the whole set of all six domains.

In candidate improvement criteria where acute phase reactants were omitted—resulting in criteria with five

domains—similar results were obtained to the corresponding criteria where acute phase reactants were included (table 3). This shows that inclusion or exclusion of these variables had no influence on the performance of the improvement criteria. Candidate improvement criteria included the domain *spinal mobility*, as assessed by a single measure instead of the whole metrology index (the BASMI), which is a combination of five mobility measures. Standardised response means (SRM) were calculated using the data from the infliximab trial (comparing baseline with week 12) for all five measures of the BASMI. It should be mentioned that all five BASMI measures are semiquantitative; their ranges, in centimetres or degrees scored as 0, 1, or 2, were used for calculations of the SRM values. This indicates that a change of, for example, 1 cm or 10% from 10 to 11 for the lateral lumbar flexion corresponds to a change from 0 to 1 on the semiquantitative scale, which would be a 100% change. Linear scales for the five measures would be better but were not available in the trial datasets and therefore could not be used.

Lateral lumbar flexion, with an SRM of 0.84, showed the best representation of the total BASMI, followed by anterior lumbar flexion (Schober test) with an SRM of 0.64. All three other measures (cervical rotation, tragus to wall distance, and intermalleolar distance) worked less well, with low SRMs. We therefore tested improvement criteria in which spinal mobility was assessed with this single mobility measure alone. The data show that the χ^2 values were as good as those obtained with the whole BASMI (table 3).

In both trials, about 35–50% of patients had normal values at baseline (= 0) and another 30–40% had a flexion of 5–10 cm (= 1). This indicates that almost half the patients could not improve because of initially normal measurements.

When we tested improvement criteria already used in recent trials with anti-TNFα agents, the highest χ^2 value of 23.5 was found for a ⩾40% change in BASDAI (table 4). This value was somewhat lower than, but not significantly different from, reference improvement criterion 1 (table 3). With a cut off of 40%, the ASAS response criteria showed similar results, as indicated by a χ^2 value of 26.5 (table 4).

Validation of the candidate criteria using the second combined dataset from the etanercept trial showed that most of the best performing candidate criteria from the infliximab dataset also had the highest χ^2 values in the second combined dataset (tables 2 to 4), indicating good reliability of the set criteria.

DISCUSSION

As anti-TNF agents are internationally considered to represent major progress in the treatment of AS, and as they seem to be far more effective than NSAIDs, it is apparent that clinically relevant criteria are needed that perform better than those developed for assessing NSAIDs.[16] The present analysis of data on almost 100 patients treated with the anti-TNFα agents infliximab and etanercept was likely to provide a sound basis for the development of improvement criteria for anti-TNFα therapy in AS. We therefore analysed the data collected in our randomised placebo controlled trials on infliximab[1] and etanercept[8] so that we could propose improvement criteria for anti-TNFα agents in AS. As the sample size in this study was not extensive, we compared the candidate criteria not only by their χ^2 values but also by the 95% confidence limits of the χ^2 values, and applied a statistical subset selection procedure to analyse the differences between candidate criteria in relation to the reference criterion.

On the basis of the data from both trials with anti-TNFα agents in AS, we propose the following two improvement criteria for further consideration: a ⩾20% change in five of six domains, and the 40% modification of the ASAS response criteria. As neither of these two sets is clearly superior on statistical grounds, the final decision needs to be taken by expert opinion. The performance of these two proposed improvement criteria should be further tested in other datasets from ongoing trials with both these anti-TNFα agents.

The two sets of improvement criteria are different in several aspects: one uses a cut off of 20% improvement in five of six domains, including spinal mobility and C reactive protein as more objective measures, while the other takes advantage of the already established ASAS improvement criteria but increases the cut off to 40% improvement in three of the four domains: patient's global, pain, function, and morning stiffness. In terms of simplicity, the ASAS 40 improvement criteria may have the advantage.

In contrast to treatment with NSAIDs, anti-TNFα therapy is very likely to have disease controlling properties, so assessment of improvement in this area of treatment might well include domains for assessment of function and inflammation such as spinal mobility and acute phase reactants. These domains are known not to be influenced by NSAIDs to a large extent,[16] while both are known to be influenced by anti-TNFα treatment.[1] Thus including these measures seems advantageous when mean group levels are compared. However, when looked at in more detail and in individual patients, the problem arises as to whether a 20% improvement in C reactive protein or spinal mobility is a reliable cut off point. For example, can we be sure that a patient whose C reactive protein has improved from 10 mg/dl to 8 mg/dl (a 20% improvement) is really obtaining benefit? And what about the patient with a 20% increase in lateral spinal flexion (for example from 10 cm to 12 cm)? From our RCT data we cannot answer these questions because all measurements were part of the BASMI, which means that they were less precise owing to the semiquantitative system involved. This implies that every improvement is by definition a 50% or a 100% step. Furthermore, as shown in the results section, C reactive protein could be taken out of the improvement criteria set without loss of information. Clearly, C reactive protein will not be helpful in patients with active disease but with low or normal C reactive protein levels. These examples indicate that 20% improvement, even of more ''objective'' indices, may not be the ultimate solution for an optimal set of improvement criteria. At the other hand, these are the only objective domains that have high face validity as being important in controlling the disease process. Although 20% appears to be a small improvement, it is difficult to achieve a consistent improvement across at least five different domains. Moreover, the available data are from six to 12 week trials. So if an improvement of spinal mobility of this magnitude can be achieved in this short treatment period it is rather impressive and has not been matched by any other treatment. Clearly, more data on the impact on long term outcome need to be established.

The ASAS improvement criteria cover only improvements in signs and symptoms of the disease, because other measures (including spinal mobility) are not sensitive to change in AS patients treated with NSAIDs. All four domains of the ASAS improvement criteria are assessed by patient questionnaires which are subjective by their nature. The original ASAS 20% improvement criteria, as developed in and for trials with NSAIDs, had a rather low sensitivity (50%) but a higher specificity (75%) for showing improvement in AS patients.[16] These values indicate that only 50% of the responders to NSAIDs are detected by these criteria but also that 25% of the non-responders to NSAIDs are falsely identified as responders. In a subsequent paper, patients and physicians considered that the response rate is underestimated by using these criteria, but that patients who are

judged to be responders according to the criteria are indeed responders.[22] However, among a whole range of possible criteria these ones clearly performed best.[16 22]

As shown in the results section, the situation of the ASAS 40% improvement criteria when used in anti-TNFα trials is clearly different, as more than 60% of the responders are identified and only 5% of responders were found in patients who had received placebo treatment. However, it should be emphasised that response rates obtained in NSAID trials cannot be compared directly with those obtained in anti-TNFα trials. NSAID trials use a so called ''flare design'': patients already use NSAIDs before inclusion in the trial; they have to stop treatment; and only if there is a certain increase in symptoms (for example, 30%) and a certain level of symptoms (for example, >4) will they be included in the trial. Thus the patients are already known NSAIDs responders. In contrast, in anti-TNFα trials patients have high disease activity despite the use of NSAIDs,

Nonetheless, considering the differences between NSAID and anti-TNFα trials, anti-TNFα therapy has a much greater efficacy than NSAIDs. This has been consistent in all trials published so far[1–7] in which high percentages of patients showed an improvement in disease activity of more than 50% in comparison to baseline. Accordingly, in most clinical studies on the efficacy of anti-TNFα therapy in AS, a 50% improvement in the BASDAI has been used as one of the main outcome variables.[1] The ASAS experts have recently recommended that the efficacy of anti-TNFα therapy should be monitored by measuring BASDAI in clinical practice and, as already mentioned, that consideration should be given to discontinuing it in patients whose response is less than 50%. Although the BASDAI 50% improvement criterion performed less well in this dataset than the two improvement criteria discussed above, it seems possible for it to be used in clinical practice because it is well known and easy to perform.

One important additional aspect of this study was the finding that the lateral spinal flexion test alone seems to represent the total BASMI which comprises four other measures. This suggests that it might be sufficient to use this tool to assess spinal mobility in AS trials with anti-TNFα agents. However, it needs to be stressed that the BASMI does not include continuous measures but subscales (for example >20 cm = 0, 10–20 cm = 1, <10 cm = 2). Furthermore, chest expansion was not included in these analyses, a variable that has also been found to be sensitive to change during anti-TNFα trials.[5]

Conclusions

In all, this analysis showed that two sets of improvement criteria are useful for assessing improvement in AS trials. No final decision about which set to use was taken after an expert session of the ASAS working group in October 2002. There will be further possibilities to study the performance of these criteria in the large clinical trials of anti-TNF treatment which are now ongoing. Lateral spinal flexion seems to be a useful instrument for measuring spinal mobility in AS patients.

Authors' affiliations

J Brandt, J Sieper, M Rudwaleit, Department of Gastroenterology/Rheumatology, Charité, Campus Benjamin Franklin, Berlin, Germany
J Listing, German Rheumatism Research Centre, Berlin, Germany
J Braun, Rheumatology Centre Ruhrgebiet, Herne, Germany
D van der Heijde, University Hospital Maastricht, Maastricht, Netherlands

REFERENCES

1 **Braun J**, Brandt J, Listing J, Zink A, Alten R, Krause A, *et al*. Treatment of active ankylosing spondylitis with infliximab, a double-blind placebo controlled multicenter trial. *Lancet* 2002;**359**:1187–93.
2 **Van Den Bosch F**, Kruithof E, Baeten D, Herssens A, de Keyser F, Mielants H, *et al*. Randomized double-blind comparison of chimeric monoclonal antibody to tumor necrosis factor alpha (infliximab) versus placebo in active spondylarthropathy. *Arthritis Rheum* 2002;**46**:755–65.
3 **Stone M**, Salonen D, Lax M, Payne U, Lapp V, Inman R. Clinical and imaging correlates of response to treatment with infliximab in patients with ankylosing spondylitis. *J Rheumatol* 2001;**28**:1605–14.
4 **Maksymowych WP**, Jhangri GS, Lambert RG, Mallon C, Buenviaje H, Pedrycz E, *et al*. Infliximab in ankylosing spondylitis: a prospective observational inception cohort analysis of efficacy and safety. *J Rheumatol* 2002;**29**:959–65.
5 **Breban M**, Vignon E, Claudepierre P, Devauchelle V, Wendling D, Lespessailles E, *et al*. Efficacy of infliximab in refractory ankylosing spondylitis: results of a six-month open-label study. *Rheumatology* 2002;**41**:1280–5.
6 **Marzo-Ortega H**, McGonagle D, O'Connor P, Emery P. Efficacy of etanercept in the treatment of the entheseal pathology in resistant spondylarthropathy: a clinical and magnetic resonance imaging study. *Arthritis Rheum* 2001;**44**:2112–17.
7 **Gorman JD**, Sack KE, Davis JC. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor alpha. *N Engl J Med* 2002;**346**:1349–56.
8 **Brandt J**, Khariouzov A, Listing J, Haibel H, Sörensen H, Grassnickel L, *et al*. Six months results of a double-blind placebo controlled clinical trial of etanercept in active ankylosing spondylitis. *Arthritis Rheum* 2003;**48**:1667–75.
9 **Braun J**, Sieper J. Therapy of ankylosing spondylitis and other spondyloarthritides: established medical treatment, anti-TNF-alpha therapy and other novel approaches. *Arthritis Res* 2002;**4**:307–21.
10 **Brandt J**, Haibel H, Cornely D, Golder W, Gonzalez J, Reddig J, *et al*. Successful treatment of active ankylosing spondylitis with the anti-tumor necrosis factor alpha monoclonal antibody infliximab. *Arthritis Rheum* 2000;**43**:1346–52.
11 **Brandt J**, Haibel H, Reddig J, Sieper J, Braun J. Treatment of patients with severe ankylosing spondylitis with infliximab – a one year follow up. *Arthritis Rheum* 2001;**44**:2936–37.
12 **Garrett S**, Jenkinson TR, Kennedy LG, Whitelock HC, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis. The Bath AS disease activity index. *J Rheumatol* 1994;**21**:2286–9.
13 **Felson DT**, Anderson JJ, Boers M, Bombardier C, Furst D, Goldsmith C, *et al*. American College of Rheumatology. Preliminary definition of improvement in rheumatoid arthritis. *Arthritis Rheum* 1995;**38**:727–35.
14 **Felson DT**, Anderson JJ, Lange ML, Wells G, LaValley MP. Should improvement in rheumatoid arthritis clinical trials be defined as fifty percent or seventy percent improvement in core set measures, rather than twenty percent? *Arthritis Rheum* 1998;**41**:1564–70.
15 **Braun J**, Pham T, Sieper J, Davis J, van der Linden S, Dougados M, *et al*, for the ASAS working group. International ASAS consensus statement for the use of anti-tumour necrosis factor agents in patients with ankylosing spondylitis. *Ann Rheum Dis* 2003;**62**:793–4.
16 **Anderson JJ**, Baron G, van der Heijde D, Felson DT, Felson M. ASAS preliminary criteria for short term improvement in ankylosing spondylitis. *Arthritis Rheum* 2001;**44**:1876–86.
17 **Van der Heijde D**, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden S. Preliminary core sets for endpoints in ankylosing spondylitis. *J Rheumatol* 1997;**24**:2225–9.
18 **Van der Heijde D**, Calin A, Dougados M, Khan MA, van der Linden S, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy, and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. *J Rheumatol* 1999;**26**:951–4.
19 **Calin A**, Garrett S, Whitelock HC, Kennedy LG, O'Hea J, Mallorie P, *et al*. A new approach to defining functional ability in ankylosing spondylitis. The Bath AS functional index. *J Rheumatol* 1994;**21**:2286–91.
20 **Jenkinson TR**, Mallorie PA, Whitelock HC, Kennedy LG, Garrett S, Calin A. Defining spinal mobility in ankylosing spondylitis. The Bath AS metrology index. *J Rheumatol* 1994;**21**:1694–8.
21 **Spoorenberg A**, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, *et al*. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. *J Rheumatol* 1999;**26**:980–4.
22 **van Tubergen A**, van der Heijde D, Anderson J, Landewe R, Dougados M, Braun J, *et al*. Comparison of statistically derived ASAS improvement criteria for ankylosing spondylitis with clinically relevant improvement according to an expert panel. *Ann Rheum Dis* 2003;**62**:215–21.